Identifying Insects with Incomplete DNA Barcode Libraries, African Fruit Flies (Diptera: Tephritidae) as a Test Case

M. Virgilio\textsuperscript{1,2}, K. Jordaens\textsuperscript{1,3}, F.C. Breman\textsuperscript{1,3}, T. Backeljau\textsuperscript{2,4}, M. De Meyer\textsuperscript{1}

\textsuperscript{1} Royal Museum for Central Africa, Leuvensesteenweg 13, 3080, Tervuren, Belgium,
\textsuperscript{2} Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000, Brussels, Belgium,
\textsuperscript{3} Joint Experimental Molecular Unit, Leuvensesteenweg 13, 3080, Tervuren, Belgium,
\textsuperscript{4} Department of Biology, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium
DNA barcoding:

molecular identification of a species based on the reference sequence with the lowest genetic distance

(Ratnasingham and Hebert 2007)
the Best Match criterion (BM)

query assigned the species name of its best-matching barcode regardless of how similar the query and barcode sequences are

<table>
<thead>
<tr>
<th>unknown query</th>
<th>reference database</th>
<th>genetic distances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>list of the more closely related sequences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>closest matches</th>
<th>species</th>
<th>genetic distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>barcode a</td>
<td>sp.1</td>
<td>2.0%</td>
</tr>
<tr>
<td>barcode b</td>
<td>sp.1</td>
<td>2.5%</td>
</tr>
<tr>
<td>barcode c</td>
<td>sp.2</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

ID= sp.1
an ideal scenario:

100% taxon coverage +
clear separation between
intra- and inter-specific variability

unknown query (sp. 5)

correct match!

complete reference library of DNA barcodes

sp. 1
sp. 2
sp. 3
sp. 4
sp. 5
sp. 6
sp. 7
sp. 8
sp. 9
sp. 10
...

a more realistic scenario:

barcode library with incomplete taxon coverage

unknown query (sp. 5)

wrong match!

misidentification

incomplete reference library of DNA barcodes

sp. 1
sp. 2
sp. 3
sp. 4
sp. 5 missing
sp. 6 missing
sp. 7 missing
sp. 8
sp. 9
sp. 10
...

incomplete taxon coverage
A RULE of THUMB:

queries with suspiciously HIGH GENETIC DISTANCES with their best match might be misidentified and suggest that there might be NO CONSPECIFIC reference sequences for that query in the library

library of 602 tephritid DNA barcodes

![Genetic Distances (K2P) Distribution](chart.png)

- **K2P > 3%**
- **2% < K2P < 3%**
- **1% < K2P < 2%**
- **K2P < 1%**
a DISTANCE THRESHOLD might help reducing the misidentification of unrepresented queries

the Best Close Match criterion (BCM)

1. establish a distance threshold:
2. accept the ID when distance query-best match < threshold
3. reject " " " " distance query-best match > threshold
BCM possible outcomes:

- **ID accepted**
  - (1) true positive $Q_x = SpA$
  - (2) false positive $Q_x \neq SpA$

- **ID rejected**
  - (3) true negative $Q_x \neq SpA$
  - (4) false negative $Q_x = SpA$

$Q_x =$ unknown query
$SpA =$ species A

Qx = SpA
several distance thresholds have been proposed:

- 10x threshold (Hebert et al. 2004)
- marine gastropods: \(3.2 \times \) to \(6.8 \times\) (Meyer and Paulay 2005)
- Lefevure et al. (2006)
- BOLD is now using an arbitrary 1\% distance threshold

but no universal threshold is applicable to all taxonomic groups
insect DNA barcodes in BOLD (25 nov 2011)

barcoded species ≈ 11% described species
insect DNA barcodes in BOLD  
(25 nov 2011)

taxonomic bias
two big questions in DNA barcoding:

• how to define a reliable distance threshold?
• how to use an incomplete reference library?

objectives of this study:

• build a reference library for African tephritids
• estimate a theoretical distance threshold producing a known ID error (e.g. < 5%)
• test if it might work in the real world
the reference library for African tephritids:

- 602 DNA barcodes, 154 species
- 33 pests + 121 non pests

DNA barcodes from
- 30 African countries (89.5%)
- adjacent islands and archipelagos (5.6%)
- other non-African countries (4.9%)

<table>
<thead>
<tr>
<th>species</th>
<th>species</th>
<th>barcodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bactrocera</td>
<td>9</td>
<td>84</td>
</tr>
<tr>
<td>Bistrispinaria</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Capparimyia</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Carpophthoromyia</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Ceratitis</td>
<td>53</td>
<td>276</td>
</tr>
<tr>
<td>Clinotaenia</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Dacus</td>
<td>61</td>
<td>187</td>
</tr>
<tr>
<td>Neoceratitise</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Perilampsis</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Trirhithrum</td>
<td>14</td>
<td>28</td>
</tr>
</tbody>
</table>
simulation of tephritid barcoding

each reference DNA barcode in the library
used as a query against all the others

30 arbitrary distance thresholds:
- TP, TN, FP, FN
- accuracy
- precision
- overall ID error
- relative ID error
Tephritid library
(602 barcodes)

- true positives (TP)
- false positives (FP)
- true negatives (TN)
- false negatives (FN)
- accuracy: (TP+TN)/n.queries
- precision: TP/(TP+FP)
- discarded queries

BCM distance threshold (K2P)
relative ID error
FP / n. of not discarded queries

ESTIMATED
(from 602 reference DNA barcodes)

\[ \text{THR}_{K2P_{0.05}} = 0.011 \]
relative ID error
FP / n. of not discarded queries

ESTIMATED
(from 602 reference DNA barcodes)

\[ \text{THR}_{K2P_{0.05}} = 0.011 \]

OBSERVED
(in 188 independent tephritid queries)

relative ID error = 0.026

---

tephritid library

![Graph showing estimated and observed relative ID error](image)

estimated relative ID error

observed relative ID error

BCM distance threshold (K2P)
simulations:

• larger reference libraries (13,914 insect DNA barcodes)

• relationships between taxon coverage and DNA barcoding performance
ID errors vs. taxon coverage

overall ID error:
\[(FP + FN) / \text{tot. n. of queries}\]

relative ID error:
\[FP / \text{n. of not discarded queries}\]

Lepidoptera
7577 DNA barcodes, 1168 species

100% taxon coverage
1160 species
7572 barcodes

75% taxon coverage

50% taxon coverage

25% taxon coverage

BCM distance threshold (K2P)
ID errors vs. taxon coverage

overall ID error:
\(\frac{FP + FN}{\text{tot. n. of queries}}\)

relative ID error:
\(\frac{FP}{\text{n. of not discarded queries}}\)

Hymenoptera

2067 DNA barcodes, 160 species

<table>
<thead>
<tr>
<th>Taxon Coverage</th>
<th>Overall ID Error</th>
<th>Relative ID Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td></td>
<td>THR_{0.05} = 0.256 +/- 0.005</td>
</tr>
<tr>
<td>75%</td>
<td></td>
<td>THR_{0.05} = 0.100 +/- 0.002</td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td>THR_{0.05} = 0.048 +/- 0.001</td>
</tr>
<tr>
<td>25%</td>
<td></td>
<td>THR_{0.05} = 0.020 +/- 0.001</td>
</tr>
</tbody>
</table>

BCM distance threshold (K2P)
ID errors vs. taxon coverage

overall ID error:
\[(FP + FN) / \text{tot. n. of queries}\]

relative ID error:
\[FP / \text{n. of not discarded queries}\]
CONCLUSIONS

The threshold we propose ($\text{THR}_{K2P_{0.05}}$) can be easily estimated from the available reference libraries and should be used as a cutoff mark defining whether:

(a) we can proceed identifying the query with a known estimated statistical error (e.g. $< 0.05$)

(b) we have to discard the query and consider alternative/complementary identification methods
Acknowledgements

We would like to thank the “Joint Experimental Molecular Unit (JEMU-RBINS/RMCA)” for support in the laboratory analyses, Valérie Balmès (Laboratoire National de la Protection des Végétation, France), Sharon Reid (Food and Environment Research Agency, UK), Komina Amevoin (Université de Lomé, Togo) and Mervyn Mansell (USDA/APHIS Pretoria Branch, South Africa) for providing samples from interceptions and Norman Barr (APHIS – USA) for providing part of Ceratitis DNA barcodes used in the tephritid library. MV was funded by the Belgian Science Policy (BELSPO) Action 1 project MO/37/029.